arXiv:2603.01042v1 [cs.CL] 1 Mar 2026

Thoth: 中期训练使大语言模型具备时间序列理解能力

Jiafeng Lin [*1] Yuxuan Wang [*1] Jialong Wu [*1] Huakun Luo [1] Zhongyi Pei [1] Jianmin Wang [1]

摘要

大语言模型（LLM）在通用推理方面取得了显著成功。然而，它们仍难以理解和推理时间序列数据，这限制了其在依赖时间动态的决策场景中的效用。本文提出**Thoth**，首个具备通用时间序列理解能力的中期训练大语言模型系列。作为关键的中间阶段，中期训练实现了时间序列与自然语言之间不依赖任务和领域的对齐。为此，我们构建了*Book-of-Thoth*，一个高质量、以时间序列为中心的中期训练语料库。*Book-of-Thoth*支持时间序列到文本和文本到时间序列的生成，使大语言模型具备对时间模式的基础理解。为了更好地评估高级推理能力，我们进一步提出了 KnoTS，这是一个用于知识密集型时间序列理解的新型基准测试，旨在对时间模式和领域知识进行联合推理。大量实验表明，使用*Book-of-Thoth* 进行中期训练后，Thoth 在一系列时间序列问答基准测试中显著优于其基础模型和先进的大语言模型。此外，在数据稀缺条件下进行微调时，Thoth 也展现出更强的能力，凸显了中期训练对时间序列理解的有效性。代码见：https://github.com/thuml/Thoth。

1. 引言

大语言模型（LLM）（Achiam 等人，2023；Zhao 等人，2023）已展现出卓越的能力，能够进行序列建模、复杂推理（Wei 等人，2022）和数学问题求解（Azerbayev 等人，2023），这在很大程度上得益于对海量文本语料进行大规模预训练。然而，现实世界的数据并非都以语言形式呈现，而往往被记录为时间序列。这类时间序列数据支撑着众多实际应用，涵盖金融（Tsay，2005）、医疗保健（Kaushik 等人，2020年）和交通运输（Lippi 等人，2013）。因此，将大语言模型与时间序列数据相结合，是使基于大语言模型的系统能够对具有时间背景的信息进行推理，并在现实场景中支持更可靠决策的关键一步。

尽管前景可期，大语言模型从根本上建立在语言建模之上，并非专为建模时间序列数据而设计，因此难以捕捉细粒度的时间依赖关系以及时间序列数据固有的复杂动态。为弥补这一不足，近期研究主要依赖监督微调（SFT），利用特定任务的标注数据集使大语言模型适应时间序列任务（Wang 等人，2025a；Kong 等人，2025年）。尽管此类方法在特定任务场景中效果显著，但往往需要大量标注数据，且在不同时间模式、任务和领域之间的泛化能力有限，从而限制了其可扩展性，也凸显了后训练的局限性。

近年来，大语言模型研究的进展显著推动了研究重点向中期训练转移。中期训练位于大规模预训练与特定任务后训练之间，作为一个中间阶段，旨在调和通用能力与领域专门化之间的关系。如图所示1，它通过纳入大规模、高质量的以时间序列为中心的语料库，扩展模型的知识覆盖范围，从而发挥桥梁作用。通过保留一定比例的通用数据，这一阶段既能维持大语言模型的基础能力，又能同时强化目标能力。这种中间适配使模型能够更好地为后续后训练做好准备，从而更稳定、更有效地过渡到专门任务对齐。

尽管领域扩展取得了越来越大的成功，但时间序列分析中的中期训练仍鲜有探索，主要原因是它高度依赖合适的大规模训练数据。与面向特定任务的微调不同，中期

[1] 清华大学软件学院、BNRist。Jiafeng Lin <lin-jf21@mails.tsinghua.edu.cn>。Yuxuan Wang<wangyuxu22@mails.tsinghua.edu.cn>，Jialong Wu<wujialong0229@gmail.com>。通讯作者：Zhongyi Pei <peizhyi@tsinghua.edu.cn>。

初步工作。

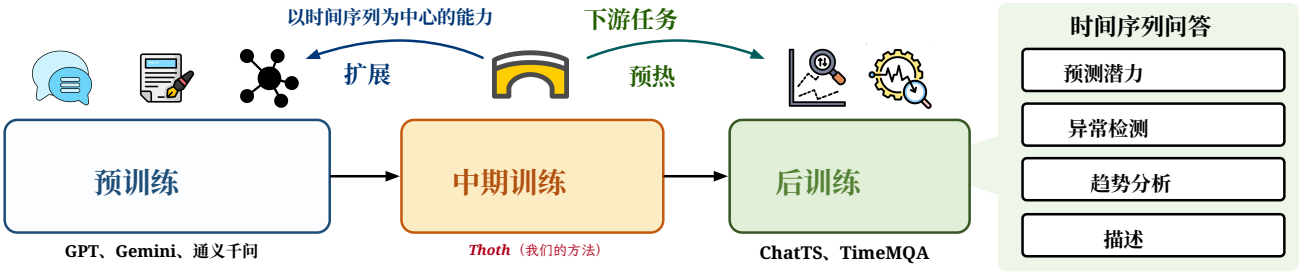


图*1*。 中期训练在连接大语言模型与时间序列理解方面发挥的关键作用。中期训练拓展了预训练大语言模型的能力，同时为后续针对特定时间序列问答任务的后训练提供关键过渡和预热。

训练旨在赋予模型通用能力，这要求训练数据涵盖范围广、结构多样，并以语言为基础。然而，大多数真实世界的时间序列多模态数据集都狭隘地聚焦于单项任务，例如预测或分类；这些数据集经过整理，旨在将时间序列直接映射到下游任务标签，因此并不适合赋予大语言模型通用的时间序列理解能力。

基于这些动机，我们提出了 **Thoth** [1]，这是首个将中期训练作为中间阶段、用于连接大语言模型与时间序列理解的模型系列。为支持时间数据与自然语言之间不受任务和领域限制的对齐，我们构建了大规模、以时间序列为中心的中期训练语料库*Book-of-Thoth*。*Book-of-Thoth* 将互补的时间序列到文本和文本到时间序列生成作为中期训练的核心任务，使模型接触多样的时间动态及其语言描述。在*Book-of-Thoth* 上进行中期训练，使 Thoth 能够习得可迁移的时间表征，同时保留基础大语言模型的通用推理能力。为更好地评估大语言模型在时间序列数据及其固有领域知识方面的高级推理能力，我们推出了 KnoTS，这是一个全新的、知识密集型时间序列理解基准测试。实验结果表明，Thoth 在 KnoTS 及范围广泛的时间序列问答基准测试上均取得显著提升，展现出强大的少样本和微调能力。这些结果表明，Thoth 是通用大语言模型迈向高级时间序列理解的关键中间步骤。

我们的贡献总结如下：

- 我们提出，中期训练是赋予大语言模型时间序列理解能力的关键阶段。
- 我们自动构建了 *Book-of-Thoth*，这是一个大规模、以时间序列为中心的语料库，用于对齐时间数据与自然语言。其核心任务包括时间序列到文本生成和文本到时间序列生成。
- 我们推出了Thoth，这是首个具备通用时间序列理解能力的中期训练大语言模型系列，在多样化的时间序列问答任务中展现出强大的泛化能力。
- 我们推出了KnoTS，这是一个面向知识密集型时间序列理解的新型基准测试。

2. 相关工作

2.1. 大语言模型的中期训练

中期训练已成为大语言模型的一种重要范式。它介于大规模预训练与任务特定的后训练之间，使模型接触高质量的专业语料库，从而在保留其通用能力的同时实现领域扩展。

中期训练的有效性在广泛的领域范围内日益得到认可。在医学和法律领域，诸如 PMC-LLaMA（Wu 等人，2024）和 SaulLM（Colombo 等人，2024）等模型表明，大规模纳入领域特定文献如何弥合通用语言逻辑与高度结构化专业知识之间的差距。在代码智能方面，诸如 Codex（Chen，2021）和 DeepSeek-Coder（Guo 等人，2024）等模型利用异构编程语料库上的多阶段训练来增强复杂问题求解能力。同样，科学领域的专业数据集，例如 GeoGalactica（Lin 等人，2023）所使用的数据集，提升了细分科学领域中的技术推理。

然而，在这些知识孤岛中进行密集训练有时会导致“垂直遗忘”（Tu 等人，2025年），使模型丧失通用能力。以往研究探索了基于正则化和基于复习的方法，以缓解这一问题。

[1]Thoth是古埃及的智慧与月亮之神，传统上被描绘为拥有朱鹭独特的头部。他被认为创造了文字、计算了时间和季节，并发明了历法。许多古埃及文献都被归于他名下，统称为《Thoth之书》。

2.2. 时间序列语言模型

时间序列建模与大语言模型的交叉领域近来受到越来越多的关注，尤其是在时间序列问答（时间序列问答）的背景下，推动了多样化数据集和专门建模技术的发展。最初，Merrill 等人（2024年）通过人类专家评估表明，即使是能力很强的大语言模型，其时间序列理解能力也十分有限。ChatTS（Xie 等人，2024年）将时间序列视为一种原生模态，并利用合成数据进行对齐，从而建立了一个多模态框架。ChatTime（Wang 等人，2025a）创新性地将时间数据建模为一种“外语”，通过多阶段训练实现零样本时间序列预测和双模态交互。为扩展任务多样性，Time-MQA（Kong 等人，2025年）引入了时间序列问答数据集，将多个时间序列任务统一到一个综合性的多任务问答框架中。ITFormer（Wang 等人，2025b）进一步通过轻量级指令微调范式连接时间编码器与冻结的大语言模型，推进了高效集成。最近，TIMEOMNI-1（Guan 等人，2025年）引入 TSR-SUITE 以纳入思维链（思维链）推理，将关注点从表层对齐转向复杂的因果理解与决策。

尽管取得了进展，现有的大多数方法仍严重依赖于在特定任务数据集上进行指令微调。通用大语言模型的最新进展表明，在大规模预训练与任务级适配之间进行中期训练，通常有助于提升模型在特定领域的表现（Tu 等人，2025年）。然而，时间序列领域的这一中间阶段仍未得到充分探索，因此有必要针对时间序列理解设置中期训练阶段。

3. 问题形式化

时间序列问答是将大语言模型与时间序列数据相结合的一种常见范式。给定单变量时间序列数据 $\mathbf{x} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots\ldots, \mathbf{x}_T\} \in \mathbb{R}^T$ 以及自然语言查询 $\mathcal{Q}$，目标是生成文本形式的回答 $\mathcal{A}$，该回答 $\mathcal{Q}$ 需以 $\mathbf{x}$ 中包含的证据为依据。形式化地说，该任务可表述如下：

$$f_\theta(\mathcal{Q}, \mathbf{x}) \to \mathcal{A}, \tag{1}$$

其中，f_θ 表示对查询 $\mathcal{Q}$ 和时间序列 $\mathbf{x}$ 进行建模，以生成相应答案 $\mathcal{A}$ 的模型。根据查询内容，回答可能是直接根据观测值得出的描述，也可能是更高层次的上下文感知推理，将 $\mathcal{Q}$ 中提供的语义背景纳入考量，并使用 $\mathbf{x}$ 中的证据论证结论。

4. 方法

4.1. Thoth之书：中期训练语料库

我们推出了*Book-of-Thoth*，这是一个全面的数据集，专为大语言模型的中期训练而设计，融合了时间序列数据与自然语言描述。与现有数据集相比，后者通常针对预测或模式分类等定义狭窄的目标进行设计，*Book-of-Thoth* 旨在捕捉时间模式与文本表示之间的双向关系，为时间序列理解提供坚实的基础，如图2所示。

具体而言，*Book-of-Thoth*由两个相互关联的部分组成：(1) 时间序列到文本，即为每个时间序列添加自然语言说明，突出其显著的时间特征；(2) 文本到时间序列，即将文本描述与相应的时间序列对齐。为确保可扩展性和一致性，我们开发了一套自动化流程，用于生成这两类高质量的时间序列-文本对。

时间序列到文本 该流程从时间序列生成开始。我们采用 KernelSynth（Ansari 等人，2024），这是一种基于高斯过程的方法，可合成具有丰富多样时间模式的时间序列。这种领域无关的合成生成，为大规模数据构建提供了稳定且可控的基础。随后，我们通过精心设计的提示词，使用 GPT-5.2 从多个角度生成自然语言描述。生成的文本描述大致可分为结构化和非结构化两类。结构化描述从特定角度对时间序列属性进行详细且高质量的刻画，例如趋势、季节性和峰值行为。相比之下，非结构化描述由简洁的说明文字或摘要构成，着重突出最显著的时间特征。这两种形式的文本数据互为补充，使模型能够同时学习细粒度的时间模式和高层次的语义抽象。

文本到时间序列 基于时间序列到文本阶段获得的结构化描述，我们通过反转时间序列与结构化文本数据，构建文本到时间序列对。每条结构化描述都明确指出时间特征，并与上一阶段生成的相应时间序列配对。这种构建方式不会引入额外的综合分析，并能保持文本与时间序列之间的一致对齐。

文本化时间序列知识 此外，我们还纳入了两部权威参考资料中的开源时间序列知识：《预测：原理与实践》(Hyndman 和 Athanasopoulos，2018）以及《时间序列分析及其应用》(Shumway，2006）。这些资料

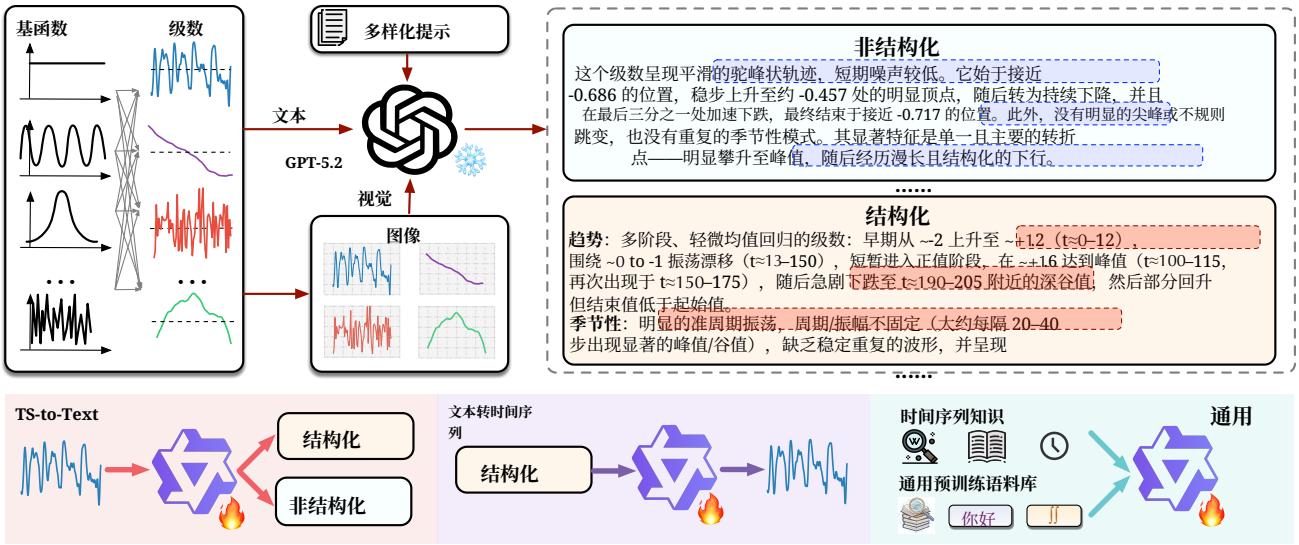


图2. 该自动化流程首先通过多种基函数的组合来合成时间序列。对于每条生成的序列，原始数值及其可视化结果都会提供给 GPT-5.2，以生成多样化的时间序列到文本描述。随后，这些文本描述与相应的时间序列配对，从而构建互补的文本到时间序列生成数据。此外，还纳入了一份经过整理的文本化时间序列知识语料，以便在中期训练期间保持预训练模型的通用能力。

为时间序列分析与预测提供了坚实的理论基础和实用指南。我们还使用 GPT-4o-mini 进行数据清理，包括移除无关内容并优化逻辑，以提高训练数据的质量和一致性。

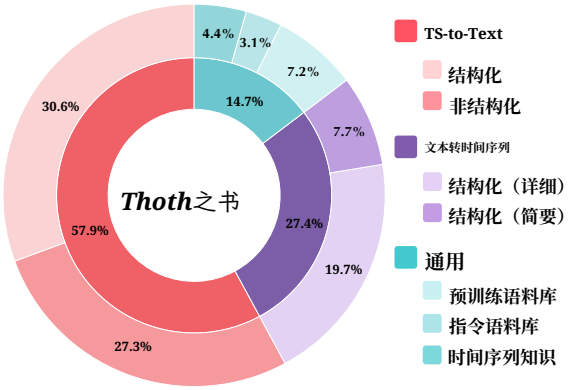


图3. *Thoth*之书 的组成部分，共计26.6M个token。

为缓解中期训练过程中的灾难性遗忘，我们采用标准的中期训练策略，将少量通用大语言模型预训练数据混入训练语料库。具体而言，我们加入了 C4（Raffel等人，2020年）和 No Robots（Rajani等人，2023）数据集中的样本。这种数据混合有助于保留通用语言与世界知识，同时有效适应时间序列专属任务。各组成部分的占比见图3。

4.2. Thoth：经过中期训练的大语言模型

架构 Thoth 是一个基于通义千问 Qwen3 架构构建的模型系列，并在两个模型变体上进行中期训练：Qwen3-30B-A3B-Instruct 和 Qwen3-8B（Yang等人，2025年）。两个变体均采用 Qwen3 的核心设计，包括用于高效序列建模的分组查询注意力（GQA）和旋转位置嵌入（RoPE），以及用于确保训练稳定性的 QK-Norm 和 RMSNorm。

训练 我们遵循默认的Qwen3训练方案，并保持骨干架构不变。所有模型均通过全参数微调进行训练，并采用 DeepSpeed ZeRO-3以实现内存高效优化。Thoth使用 AdamW进行优化，参数为(β_1, β_2) =(0.9,0.999)，并采用余弦学习率调度。峰值学习率设为 1×10^{-5}，预热比例为0.1。训练以bfloat16精度进行，最大序列长度为4096个token。我们使用32的全局批次大小，训练共进行3个epoch，约产生1,600个优化步骤。训练在8块配备80GB显存的NVIDIA A800 GPU上进行。

5. KnoTS

现有的时间序列问答（TSQA）基准测试（Kong 等人，2025年；Wang 等人，2025a）主要聚焦于时间理解的狭窄方面，例如趋势分析、异常值识别、短期预测和简单的判断类查询。值得注意的是，这些基准测试通常基于有限的真实世界语料构建，

或者通过模板驱动的文本与合成数据配对构建。因此，这类数据集的语言多样性较低，对通用世界知识的要求也很低，进而促使模型依赖反复出现的模板进行浅层映射。这种设计还会增加模型随后针对专门的 TSQA 任务进行微调时发生灾难性遗忘的风险。

为解决这些局限，我们推出 **KnoTS**(**Kno**wledge-intensive **T**ime **S**eries QA)，这是一个基准测试，旨在评估对时间序列变化与领域知识进行联合推理的能力。它的设计受到现实场景的启发：在这些场景中，解读时间序列数据必然依赖特定领域的背景信息。我们首先确定了一组应用领域，在这些领域中，时间序列数据对分析和决策至关重要，包括金融、医疗保健、电子商务、环境和社会领域。在每个领域中，我们提示 GPT 生成样本，每个样本都包含一段时间序列及相应的领域背景信息。随后，对于每个样本，我们使用 GPT 生成一组多样的问题—答案对，这些问题和答案均基于时间序列及相关领域知识。之后，人工标注员会审核并完善这些候选内容，以确保事实准确、时间序列与解释之间对齐合理，以及推理链充分。最终的基准测试包含 300 个经过精心筛选的问答对。图4展示了 KnoTS 中的一个简化案例。KnoTS 包含两种互补的任务类型：(1) 推理任务，侧重于结合领域知识推断时间模式；(2) 决策任务，即结合时间序列数据与通用领域知识进行预测或判断。



图 4. KnoTS 基准测试中一道推理问题的简要案例，完整案例见附录E。

6. 实验

数据集 为全面评估 Thoth 对时间序列的理解能力，我们采用了两个广受认可的基准测试：ChatTime（Wang 等人，2025a）和 Time-MQA（Kong 等人，2025年）。ChatTime 专为时间序列模式识别而设计，重点关注四种基本时间特征：趋势、波动性、季节性和异常值。所有任务均以多项选择题的形式呈现，标准答案由数据生成过程精确定义。Time-MQA 是一个面向真实场景的多任务问答基准测试，涵盖七种不同任务，包括预测、插补、分类、异常检测、判断、多项选择题和开放式问答。这些基准测试共同构成了一个全面的评估框架，使我们能够评估模型对时间模式的细粒度辨别能力，以及其在真实时间序列场景中的推理能力。

基线模型 我们将 Thoth 与 15 个先进基线模型进行对比评估，涵盖四个类别：（1）专有模型：Gemini-3-Flash-Preview、GPT-4o-mini、Grok-4.1-Fast，（2）开源大语言模型：Qwen3-235B-A22B-Instruct（Yang 等，2025）、Qwen3-30B-A3B-Instruct（Yang 等，2025）、Deepseek-R1-32B（Guo 等，2025）、Mistral-Small-24B-Instruct、LLaMA3.1-8B-Instruct、Qwen3-8B（Yang 等，2025），（3）开源视觉语言模型：Qwen3-VL-30B-A3B-Instruct（Bai 等，2025）、Mistral-Small-3.2-24B-Instruct，以及（4）任务专用时间序列语言模型：ChatTS-14B（Xie 等，2024）、Time-MQA (Llama3-8B)（Kong 等，2025）、Time-MQA (Qwen2.5-7B)（Kong 等，2025）、ChatTime-7B-Chat（Wang 等，2025a）。

评估详情 为了评估 Thoth 的时间序列理解能力，同时确保与基线模型进行公平比较，我们采用广受认可的 *n-shot* 上下文学习 协议（Brown 等人，2020年）。具体而言，对于所有模型，我们在提示中为每项任务提供 3–5 个固定示例作为演示。这一范围旨在确保上下文窗口限制与不同类别中类别代表性需求之间的最佳平衡。这些示例用于引导大语言模型遵循复杂指令，并遵守特定的响应格式，例如结构化的多项选择答案。

在评估指标方面，ChatTime 中的四个子任务均采用准确率（ACC）进行评估，因为所有任务都被设计为多项选择题。对于 Time-MQA，评估指标因任务而异：预测和插补采用对称平均绝对百分比误差（SMAPE）进行评估，该指标对真实时间序列中常见的尺度变化具有稳健性。分类、异常检测、判断和多项选择题任务均采用准确率（ACC）进行评估。对于开放式任务以及我们新引入的 KnoTS，由于没有明确的真实标签，我们采用先进的 Gemini-3-Pro-Preview 作为自动评估器，它会从多个维度进行评分，包括

表 *1*。 现有基准测试的性能比较。对于 Time-MQA 任务，我们报告的评估指标包括预测和插补任务的 SMAPE、开放式任务的大语言模型评审，以及其他任务的准确率（ACC）。对于 ChatTime基准测试，表中报告了所有输入长度下的平均结果。最佳结果以**粗体** 标出，次佳结果带下划线。

模型	ChatTime基准测试				Time-MQA 基准						
	趋势 ↑	波动性 ↑	季节性 ↑	异常值 ↑	预测 ↓	Impu. ↓	分类 ↑	异常检测 ↑	评判 ↑	多项选择题 ↑	开放式 ↑
专有模型											
Gemini-3-Flash-Preview	0.837	**0.863**	0.520	0.747	0.642	0.263	0.34	0.70	0.62	0.52	6.60
GPT-4o-mini	0.850	0.433	0.427	0.673	0.546	0.252	0.16	**0.80**	0.76	0.60	7.32
Grok-4.1-Fast	0.647	0.383	**0.620**	0.747	0.503	0.303	0.14	0.68	**0.78**	0.60	7.62
开源大语言模型											
Qwen3-235B-A22B-Instruct	0.867	0.513	0.447	0.667	0.458	**0.220**	0.36	0.72	0.72	0.64	**7.76**
Qwen3-30B-A3B-Instruct	0.780	0.467	0.337	0.443	0.531	0.426	0.02	0.66	0.42	0.56	6.62
Deepseek-R1-32B	0.533	0.450	0.447	0.623	0.554	0.361	0.22	0.52	0.56	0.48	6.14
Mistral-Small-24B-Instruct	0.850	0.513	0.357	0.627	0.593	0.276	0.18	0.36	0.42	0.46	6.54
Llama-3.1-8B-Instruct	0.260	0.350	0.317	0.383	0.587	0.356	**0.66**	0.60	0.64	0.62	4.22
Qwen3-8B	0.163	0.220	0.100	0.447	0.671	0.533	0.02	0.48	0.62	0.54	6.28
开源视觉语言模型											
Qwen3-VL-30B-A3B-Instruct	0.860	0.460	0.350	0.470	0.610	0.278	0.12	0.62	0.72	0.48	7.08
Mistral-Small-3.2-24B-Instruct	0.430	0.373	0.597	0.563	0.578	0.270	0.38	0.66	0.74	0.52	6.54
任务特定时间序列语言模型											
ChatTS-14B	-	-	-	-	1.166	-	0.24	0.44	0.58	0.46	4.18
ChatTime-7B-Chat	0.980	0.870	0.730	0.963	-	-	-	-	-	-	-
Time-MQA(LLaMA3-8B)	0.653	0.433	0.300	0.393	0.528	0.307	0.44	0.58	0.72	0.64	4.80
Time-MQA(Qwen2.5-7B)	0.297	0.370	0.343	0.390	0.480	0.251	0.28	0.64	0.66	0.46	5.06
Ours											
Thoth-8B	**0.973**	0.750	0.603	**0.753**	0.522	0.333	0.10	0.58	0.72	0.64	5.98
Thoth-30B-A3B	0.957	0.770	0.573	0.740	**0.432**	0.247	0.02	0.70	**0.78**	**0.70**	7.42

事实正确性、时序逻辑、完整性和简洁性。

6.1. 主要结果

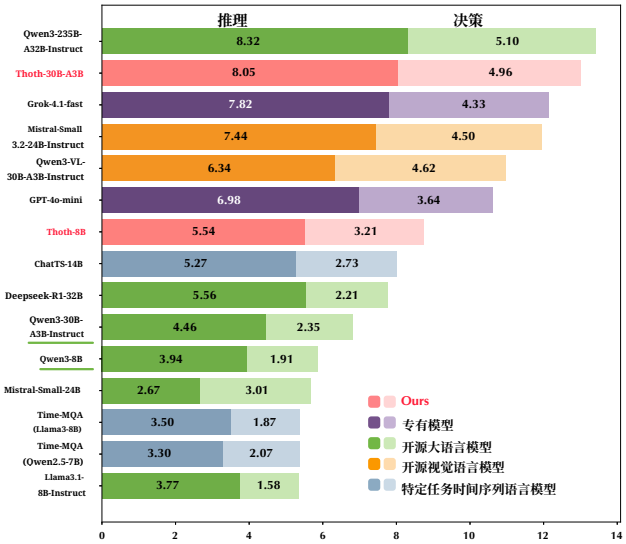


图*5*. KnoTS 基准测试中各模型在推理和决策方面的表现。所有评估均使用 Gemini-3-Pro-Preview 进行，得分范围为 0.0 至 10.0。

实验结果如表 1所示。Thoth 在时间序列问答方面展现出卓越能力。Thoth-30B-A3B 的整体表现与拥有 235B 参数的 Qwen3 模型相当，有时甚至更胜一筹。与此同时，Thoth-8B 的表现与参数量约为 30B 的开源模型不相上下，并显著优于 Qwen3-8B 基础模型。

时间序列语言模型的下游表现以灰色显示。这些模型直接在下游数据集上进行监督微调，这通常会导致明显的跨任务不平衡，以及在不同任务上的表现不稳定。值得注意的是，将时间序列图作为视觉输入的开源视觉语言模型，其表现始终优于参数量匹配的大语言模型，这凸显了视觉信息对于时间序列理解的重要性。这一观察进一步表明，使用 GPT-5.2 进行数据生成时，构建具有时间序列文本和图像双重输入的 *Book-of-Thoth* 是有效的。

ChatTime 基准测试还根据输入长度对评估结果进行了细分，完整结果见附录G。由于基于大语言模型的生成本质上具有概率性，我们还进行了误差棒分析以量化不确定性，详细结果见附录C。

除了这些公开的时间序列问答基准外，我们还在提出的 KnoTS 上开展实验，以评估其对时间序列与领域知识进行联合推理以及决策的能力。评估结果由 Gemini-3-Pro-Preview 评分，这是目前最强大的大语言模型之一。因此，Gemini-3-Flash-Preview 未被纳入基线

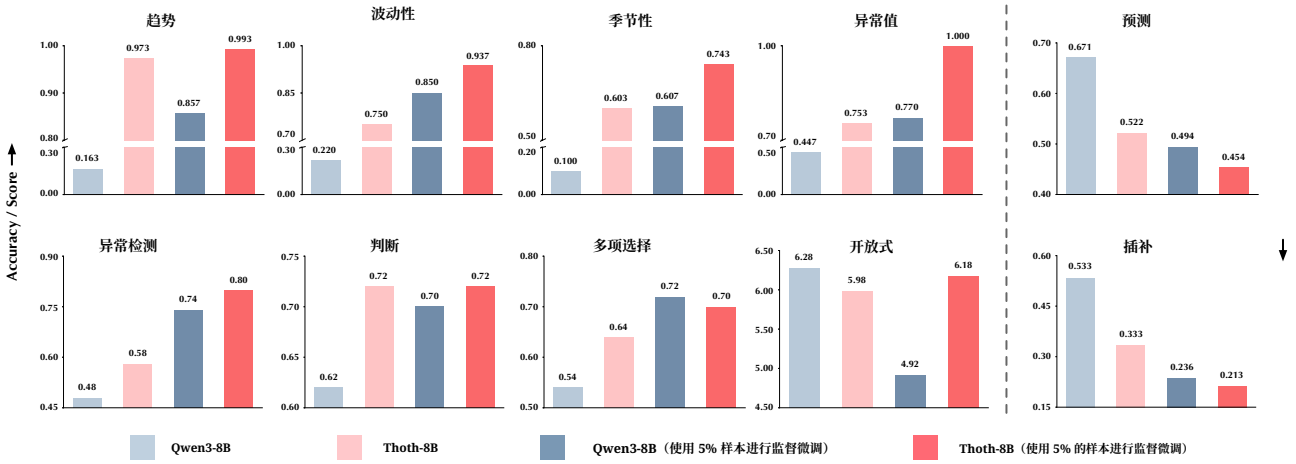


图6. 使用 5% 抽样训练集在 ChatTime 和 Time-MQA 基准测试上的监督微调性能。对于每个基准测试，均从各子任务中随机选取相同比例的数据进行训练，不会针对每个子任务单独训练模型。

模型，因为同一系列中的模型很可能共享训练语料库中的大量内容。

如图 5所示，Thoth-30B-A3B 表现出更优异的性能，超过了 Grok-4.1-Fast 等强大的大语言模型和视觉语言模型，仅略逊于总参数规模为 235B 的 Qwen3 模型。尽管 Thoth-8B 受限于其基础模型较小的参数规模，整体推理能力因此受到限制，但它通过中期训练显著增强了时间序列理解能力。因此，Thoth-8B 的表现优于大多数参数规模处于 30B 范围内的开源大语言模型，以及所有现有的时间序列语言模型。

6.2. 监督微调结果

作为一种经过中期训练的大语言模型，Thoth 旨在为后训练和特定下游任务提供关键预热。在本节中，我们分别从 Time-MQA 和 ChatTime 中随机抽取 5% 的子集对 Thoth 进行微调，同时保留各基准测试中的原始数据分布。我们将结果与使用相同数据子集直接微调原始 Qwen3-8B 模型所得的结果进行比较。如图 6 所示，经过监督微调后，Thoth 在绝大多数任务上的表现都显著优于基线大语言模型，证明了 Thoth 在下游任务预热方面的有效性。

一个有趣的观察是，在对Qwen3-8B进行微调后，开放式任务上的性能急剧下降，这表明直接在多个下游任务上进行微调可能会引发不良的能力权衡，即某些技能的提升以牺牲其他技能为代价。相比之下，经过微调的Thoth-8B实现了稳定的性能提升。这归因于其中期训练阶段，该阶段使模型能够利用其已建立的理解来吸收任务特定知识，在多任务微调过程中，而不是诉诸死记硬背。

6.3. 数据扩展

在本节中，我们通过在不同数据规模范围内开展中期训练实验，考察模型的实证扩展行为。具体而言，我们使用了若干子集

关于 *Book-of-Thoth* 的内容，分别占总量的 10%、20%、50%、75% 和 100%，同时保持数据分布一致。图7所示结果揭示了一个明显趋势：随着数据集规模增大，模型性能稳步提升，凸显了数据扩展的巨大潜力。

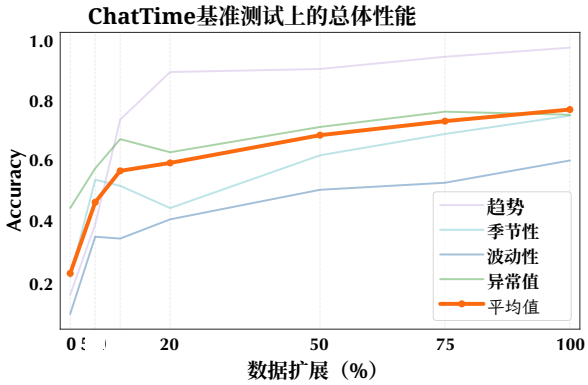


图7。 Thoth-8B 在 ChatTime 基准测试中不同中期训练数据规模下的综合表现。随着数据规模增大，Thoth-8B 的平均表现更佳。

6.4. 消融实验

为验证*Book-of-Thoth*各组件的有效性，我们通过分别移除这三个组件中的每一个，进行了详细的消融实验。结果列于表2中。移除*TS-to-Text* 组件后，由于剩余数据集主要由时间序列构成

表2. 在*Book-of-Thoth*上的消融实验结果。我们分别移除三种语料进行训练：时间序列到文本配对（TS-to-Text）、文本到时间序列配对（Text-to-TS）和通用预训练数据（通用），并将结果与 Thoth-8b 进行比较。

模型	ChatTime基准测试				Time-MQA 基准							KnoTS 基准测试	
	趋势 ↑	波动性 ↑	季节性 ↑	异常值 ↑	预测 ↓	Impu. ↓	分类 ↑	异常检测 ↑	评判 ↑	多项选择题 ↑	开放式 ↑	推理 ↑	决策 ↑
Qwen3-8B	0.163	0.220	0.100	0.447	0.671	0.533	0.02	0.48	0.62	0.54	6.28	3.94	1.91
– TS-to-Text	0.360	0.487	0.347	0.453	**0.495**	**0.286**	0.14	0.48	0.68	0.46	3.32	3.13	1.45
– 文本转时间序列	0.967	0.677	**0.630**	0.733	0.532	0.476	0.12	0.56	0.68	0.58	5.10	5.23	3.04
– 通用	0.963	0.547	0.387	**0.763**	0.538	0.363	**0.16**	**0.58**	0.70	0.60	**6.12**	5.30	3.06
Thoth-8B	**0.973**	**0.750**	0.603	0.753	0.522	0.333	0.100	**0.58**	**0.72**	**0.64**	5.98	**5.54**	**3.21**

生成任务，两个与时间序列生成相关的任务——预测和插补——性能有所提升，但整体表现仍不理想。

有趣的是，尽管文本转时间序列 提升了整体时间序列理解能力，但与*TS-to-Text*相比，提升幅度并不那么显著。这种差异可能源于这样一个事实：与根据时间序列生成文本描述相比，从文本中重建精确的时间序列要困难得多。在模型容量有限的情况下，较简单的辅助任务往往能带来更大的收益。一个典型例子是 BERT 中的掩码语言模型（MLM）任务（Devlin 等人，2019年），该任务通过捕捉词元级别的双向上下文信息，在较小模型的特征提取方面表现出很高的效率。然而，随着模型规模增大，遵循 GPT 范式、复杂度更高的下一个词元预测（NTP）任务则展现出优势，这与大模型生成能力的增强相一致（Brown 等人，2020年）。尽管目前文本转时间序列带来的提升可能不如*TS-to-Text*显著，但它仍能带来宝贵的改进，并在推动时间序列理解方面具有不可或缺的潜力。

6.5. 案例研究

本节通过图 8展示一个案例研究，呈现 Thoth-30B-A3B 在 KnoTS 基准测试上的简化推理结果。如图所示，模型基于三个不同阶段推断用户的具体行为

在时间序列分析方面。这些结果与真实值一致，展现了 Thoth 在特定领域场景中强大的时间序列推理能力。

7. 讨论

鉴于时间序列数据在现实应用中的广泛存在，赋予大语言模型时间序列理解能力变得日益重要。本文将中期训练视为使大语言模型适应时间序列理解的实用桥梁。为此，我们构建了*Thoth之书*，这是一个大规模、与任务和领域无关、以时间序列为中心的语料库，将时间序列与自然语言对齐，并用其训练了Thoth这一系列通用大语言模型

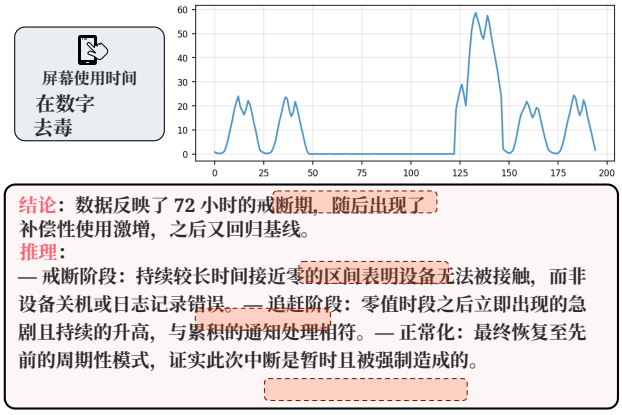


图8. Thoth-30B-A3B 模型在 KnoTS 基准测试上的推理案例研究。各模型的详细对比结果请参阅附录 F。

其时间序列理解能力得到了显著提升。为更好地评估大语言模型的时间序列推理能力，我们进一步提出了知识密集型时间序列问答基准，强调基于时间证据的推理与决策。大量实验表明，Thoth在广泛的时间序列任务上展现出出色的少样本性能。此外，监督微调研究显示，中期训练能够有效预热下游时间序列问答，使模型仅凭有限的任务专属数据也能持续获得增益。分析实验展现了出色的可扩展性，并提供了有价值的洞见。

局限性与未来工作 本文聚焦于通过中期训练将大语言模型与时间序列理解相衔接。展望未来，值得进一步探索后训练的改进，以提升时间序列理解与推理能力。图1展示了中期训练在预热下游任务方面的强大能力。然而，显然，针对下游任务直接进行多任务学习微调，可能会导致某些任务的性能下降，这是因为各任务的需求并不均衡。因此，强化学习已在以往的问答任务研究中被广泛证明有效，值得进一步探究。

表2 展示了 *Book-of-Thoth* 框架中不同组件对模型性能的影响。

文本到时间序列训练任务展现出良好的潜力；然而，正如第 6.4 节所讨论的，在更大规模的模型和数据集上仍有改进空间。此外，生成时间序列的训练信号是否应基于其与文本信息的对齐，而非传统的下一词元预测损失，仍然是一个悬而未决的问题。

参考文献

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., 等。GPT-4 技术报告。*arXiv* 预印本 *arXiv:2303.08774*，2023。

Ansari, A. F., Stella, L., Turkmen, C., Zhang, X., Mercado, P., Shen, H., Shchur, O., Rangapuram, S.S., Arango, S.P., Kapoor, S., 等。Chronos：学习时间序列语言。*arXiv* 预印本 *arXiv:2403.07815*，2024。

Azerbayev, Z., Schoelkopf, H., Paster, K., Santos, M. D., McAleer, S., Jiang, A. Q., Deng, J., Biderman, S., 和 Welleck, S。Llemma：面向数学的开放式语言模型。*arXiv* 预印本 *arXiv:2310.10631*，2023。

Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., Ge, W., Guo, Z., Huang, Q., Huang, J., Huang, F., Hui, B., Jiang, S., Li, Z., Li, M., Li, M., Li, K., Lin, Z., Lin, J., Liu, X., Liu, J., Liu, C., Liu, Y., Liu, D., Liu, S., Lu, D., Luo, R., Lv, C., Men, R., Meng, L., Ren, X., Ren, X., Song, S., Sun, Y., Tang, J., Tu, J., Wan, J., Wang, P., Wang, P., Wang, Q., Wang, Y., Xie, T., Xu, Y., Xu, H., Xu, J., Yang, Z., Yang, M., Yang, J., Yang, A., Yu, B., Zhang, F., Zhang, H., Zhang, X., Zheng, B., Zhong, H., Zhou, J., Zhou, F., Zhou, J., Zhu, Y., and Zhu, K. 通义千问3-VL 技术报告，2025年。URL https://arxiv.org/abs/2511.21631。

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., 等。语言模型是少样本学习者。神经信息处理系统进展，33：1877–1901，2020年。

Chen, J., Feng, A., Zhao, Z., Garza, J., Nurbek, G., Qin, C., Maatouk, A., Tassiulas, L., Gao, Y., 和 Ying, R。Mtbench：一个用于时间推理和问答的多模态时间序列基准测试。*arXiv* 预印本 *arXiv:2503.16858*，2025年。

Chen, M。评估基于代码训练的大语言模型。*arXiv* 预印本 *arXiv:2107.03374*，2021。

Colombo, P.、Pires, T. P.、Boudiaf, M.、Culver, D.、Melo, R.、Corro, C.、Martins, A.F.、Esposito, F.、Raposo, V. L.、Morgado, S., 等。Saullm-7b：一款开创性的法律大语言模型。*arXiv* 预印本 *arXiv:2403.03883*，2024。Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert：用于语言理解的深度双向变换器预训练。收录于*2019*年计算语言学协会北美分会会议论文集：人类语言技术，第*1*卷（长篇和短篇论文），第4171–4186页，2019年。

Guan, T., Meng, Z., Li, D., Wang, S., Yang, C.-H. H., Wen, Q., Liu, Z., Siniscalchi, S. M., Jin, M. 和 Pan, S. Timeomni-1：通过时间序列激励大语言模型进行复杂推理。*arXiv* 预印本 *arXiv:2509.24803*，2025年。

Guo, D., Zhu, Q., Yang, D., Xie, Z., Dong, K., Zhang, W., Chen, G., Bi, X., Wu, Y., Li, Y. 等。Deepseek-coder：当大语言模型遇上编程——代码智能的崛起。*arXiv* 预印本 *arXiv:2401.14196*，2024年。

Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X. 等。Deepseek-r1：通过强化学习激励大语言模型的推理能力。*arXiv* 预印本 *arXiv:2501.12948*，2025年。

Hyndman, R. J. 和 Athanasopoulos, G. 预测：原理与实践。OTexts，2018。

Kaushik, S.、Choudhury, A.、Sheron, P. K.、Dasgupta, N.、Natarajan, S.、Pickett, L. A. 和 Dutt, V. 医疗保健中的人工智能：使用统计、神经网络和集成架构进行时间序列预测。《大数据前沿》，3:4，2020年。

Kong, Y.、Yang, Y.、Hwang, Y.、Du, W.、Zohren, S.、Wang, Z.、Jin, M. 和 Wen, Q. Time-MQA：借助上下文增强进行时间序列多任务问答。*arXiv* 预印本 *arXiv:2503.01875*，2025年。

Langley, P. 撰写机器学习论文。载于 Langley, P.（编），第*17*届国际机器学习会议论文集（*ICML 2000*），第1207–1216页，美国加利福尼亚州斯坦福，2000年。Morgan Kaufmann。

Lin, Z., Deng, C., Zhou, L., Zhang, T., Xu, Y., Xu, Y., He, Z., Shi, Y., Dai, B., Song, Y., 等. GeoGalactica：地球科学领域的科学大语言模型。*arXiv* 预印本 *arXiv:2401.00434*，2023。

Lippi, M., Bertini, M., and Frasconi. 短期交通流量预测：时间序列分析与监督学习的实验比较。*IEEE* 智能交通系统汇刊，14(2):871–882，2013。

Merrill, M. A., Tan, M., Gupta, V., Hartvigsen, T., and Althoff, T. 语言模型在零样本时间序列推理方面仍面临困难。载于计算语言学协会会议论文集：*EMNLP 2024*，第 3512–3533 页，2024。Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. 探索统一文本到文本转换器的迁移学习极限。机器学习研究期刊，21 (140):1–67, 2020。Rajani, N., Tunstall, L., Beeching, E., Lambert, N., Rush, A. M., and Wolf, T. No Robots。
h
t
t
p
s://
huggingface.co/datasets/HuggingFaceH4/no_robots，2023。Shumway, R. H. 时间序列分析及其应用：R 语言实例，2006。Tsay, R. S.金融时间序列分析。约翰·威立父子出版公司，2005。Tu, C., Zhang, X., Weng, R., Li, R., Zhang, C., Bai, Y., Yan, H., Wang, J., and Cai, X. 大语言模型中期训练综述。*arXiv* 预印本 *arXiv:2510.23081*，2025。Wang, C., Qi, Q., Wang, J., Sun, H., Zhuang, Z., Wu, J., Zhang, L., and Liao, J. ChatTime：连接数值与文本数据的统一多模态时间序列基础模型。载于人工智能促进协会年会论文集，第 39 卷，第 12694–12702 页，2025a。Wang, Y., Lei, P., Song, J., Hao, Y., Chen, T., Zhang, Y., Jia, L., Li, Y., and Wei, Z. ITFormer：借助大规模多任务数据集，连接时间序列与自然语言以实现多模态问答。*arXiv* 预印本 *arXiv:2506.20093*，2025b。Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., 等。思维链提示可激发大语言模型的推理能力。神经信息处理系统进展，35:24824–24837，2022。Wu, C., Lin, W., Zhang, X., Zhang, Y., Xie, W., and Wang, Y. PMC-LLaMA：致力于构建用于医学领域的开源语言模型。美国医学信息学会杂志，31(9):1833–1843，2024。Xie, Z., Li, Z., He, X., Xu, L., Wen, X., Zhang, T., Chen, J., Shi, R., and Pei, D. ChatTS：通过合成数据对齐时间序列与大语言模型，以增强理解和推理能力。*arXiv* 预印本 *arXiv:2412.03104*，2024。Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., 等。通义千问3技术报告。*arXiv* 预印本 *arXiv:2505.09388*，2025。Zhang, Y., Zhang, Y., Zheng, M., Chen, K., Gao, C., Ge, R., Teng, S., Jelloul, A., Rao, J., Guo, X. 等。Insight miner：一个用于自然语言跨领域对齐的时间序列分析数据集。*arXiv* 预印本 *arXiv:2512.11251*，2025年。

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z. 等。大语言模型综述。*arXiv* 预印本 *arXiv:2303.18223*，1(2)，2023。

A. 下游任务

我们在两个重要的多任务时间序列问答（QA）基准测试上评估模型：Time-MQA（Kong 等人，2025年）和 ChatTime（Wang 等人，2025a）。这些基准测试涵盖了广泛的时间序列分析任务，包括传统数值分析和更高级的推理任务。数据集涉及多个领域，如医疗保健、金融、能源、交通运输等，为模型评估提供了全面的基础。以下章节将介绍这些基准测试及其相关任务的详细信息。

A.1. Time-MQA 基准

Time-MQA 框架是一个多任务时间序列问答系统，可在多项时间序列任务中整合自然语言查询。它包含约 200,000 个问答对，这些问答对源自 12 个不同领域，包括医疗保健、能源、金融、交通、物联网和网络。这一大规模数据集的设计目的，不仅是支持预测和异常检测等传统数值任务，还要通过文本增强查询处理更为复杂的推理任务。Time-MQA 的多任务结构使其成为评估模型在一系列时间序列问题上的通用性与推理能力的理想基准测试。

Time-MQA 数据集中的具体任务如下：

预测 Time-MQA 中的预测任务是根据已知的时间序列数据预测未来时间点。该任务涵盖医疗保健和能源等多个领域，要求模型根据历史观测值准确预测未来趋势。例如，根据先前的数据点预测心率趋势或能源消耗。

插补 插补任务旨在解决时间序列数据中缺失值的填补问题。时间序列数据常因传感器故障或数据收集不完整等各种原因而出现缺口。插补任务要求模型预测缺失值，从而恢复时间序列的连续性。这在医疗保健等领域尤为重要，因为数据完整性至关重要。

异常检测 异常检测对于识别时间序列数据中的异常模式或偏离至关重要。该任务要求模型检测偏离预期行为的异常值或异常事件。在工业和金融场景中，检测实时数据中的异常可以预示关键事件，例如设备故障或市场动荡。

分类 分类任务是将时间序列数据归入预定义类别。例如，来自可穿戴设备的传感器数据可被分类为不同的人类活动，如步行、坐着或慢跑。该任务检验模型识别时间序列中的模式并将其归入适当类别的能力。

判断题 判断题评估模型基于时间序列数据验证简单断言的能力。这些问题要求模型判断关于时间序列的陈述是否正确。

选择题在多项选择任务中，模型需要从一组选定选项中选出正确答案。该任务评估模型对时间序列数据的特定特征进行分类和比较的能力。

开放式问答Time-MQA 中的开放式问答任务要求模型针对时间序列数据生成详细的文本解释。这些任务不局限于简单的数值预测，还要求模型解读趋势、识别因果关系，并基于时间序列模式提供见解。

这些任务共同使 Time-MQA 成为用于评估多任务模型的综合基准测试。Time-MQA 融合了数值与文本推理，促使模型针对时间序列数据提供更深入的见解，弥合了传统数值分析与现代语言模型驱动方法之间的鸿沟。

A.2. ChatTime 基准测试

ChatTime 框架是另一种先进的多模态时间序列基础模型，可融合时间序列数据和文本信息。ChatTime 能够独特地处理双模态输入与输出，使模型能够利用

文本上下文和时间序列数据。它支持零样本时间序列预测，并可用于多模态时间序列问答任务，因此特别适合测试模型处理复杂且依赖上下文的任务的能力。

ChatTime 数据集包含专门设计的任务，用于评估模型结合上下文文本信息处理时间序列数据的能力。该数据集中的四项主要任务如下：

趋势分析 此任务涉及识别给定时间序列的整体走向。模型需要判断数据随时间呈现上升、下降还是稳定趋势。目标是识别数据点的总体变化，并据此对级数进行分类。

波动性分析 波动性分析侧重于评估时间序列中的波动幅度。模型需要识别高波动时期（数据点变化显著）或低波动时期（数据相对稳定）。此任务要求识别数据变化随时间的变化情况。

季节性分析 在季节性分析中，模型需要判断时间序列是否呈现周期性或循环性行为。这包括识别按固定间隔重复出现的模式或趋势，表明数据随时间受到季节性因素的影响。

异常值检测 异常值检测要求模型识别时间序列中显著偏离一般模式或趋势的数据点。这些异常值可能代表异常事件或异常现象，模型需要根据这些点与预期行为的偏离程度将其准确找出。

这些任务都要求模型整合时间序列数据和文本信息，不仅需要进行数值分析，还需要深入理解数据生成的背景。ChatTime 的多模态特性使其成为评估模型同时跨这两个领域进行推理能力的理想候选。

B. 评估的实现细节

我们提供评估的实现细节。对于 ChatTime基准测试，鉴于合成的时间序列问答对规模庞大，我们随机抽取每项任务中的 300 个实例作为测试集，以评估大语言模型识别时间序列模式的准确率。对于 Time-MQA 基准，我们遵循原文中的实验设置，每项任务抽取 50 个样本作为测试用例。具体而言，对于预测和插补任务，我们采用对称平均绝对百分比误差（SMAPE）作为评估指标。这样选择的理由是，数据集涵盖多个应用领域，数值方差极大，且序列平稳性各不相同；SMAPE 能够有效评估这些异质样本的生成质量。公式定义如下：

$$\text{SMAPE} = \frac{1}{n}\sum_{t=1}^{n}\frac{|F_t - A_t|}{(|A_t| + |F_t|)/2} \tag{2}$$

其中， A_t 表示实际值， F_t 表示预测值。在极少数情况下，若大语言模型生成的序列长度偏离指令要求，我们引入惩罚因子1.1。调整后的误差 E_{adj}计算如下：

$$E_{adj} = \begin{cases} E \times 1.1, & \text{if } L_{gen} \neq L_{inst} \\ E, & \text{if } L_{gen} = L_{inst} \end{cases} \tag{3}$$

其中， L_{inst} 和 L_{gen} 分别表示指令长度和生成长度。除开放式问题外，所有其他任务均以准确率作为评估指标。对于Time-MQA中的开放式任务以及KnoTS基准测试，我们使用Gemini-3-Pro-Preview进行评分。该模型综合事实正确性、时序逻辑、完整性和简洁性等多个维度给出综合得分。

C. 误差棒分析

为确保结果的可复现性和统计可靠性，我们使用不同的随机种子对主要实验进行了三次独立运行（$N = 3$）。结果以**均值** $\pm$ **标准差**的形式报告，以反映大语言模型解码和采样过程固有的随机性。这项分析证实，性能

提升在多次迭代中始终如一，并非源于有利的随机种子设定，这表明我们提出的方法具有稳定性。

表 *3.* Thoth-30B-A3B 和 Thoth-8B 在所有基准测试上的性能稳健性。结果取自三次随机运行。

模型	ChatTime基准测试				Time-MQA 基准							KnoTS 基准测试	
	趋势 ↑	波动性 ↑	季节性 ↑	异常值 ↑	预测 ↓	Impu. ↓	分类 ↑	异常检测 ↑	评判 ↑	多项选择题 ↑	开放式 ↑	推理 ↑	决策 ↑
Thoth-8B	0.973±0.003	0.750±0.005	0.603±0.007	0.753±0.003	0.522±0.019	0.333±0.013	0.10±0.00	0.58±0.01	0.72±0.02	0.64±0.02	5.98±0.17	5.54±0.24	3.21±0.18
Thoth-30B-A3B	0.957±0.003	0.770±0.010	0.573±0.010	0.740±0.007	0.432±0.021	0.247±0.014	0.02±0.00	0.70±0.00	0.78±0.01	0.72±0.02	7.42±0.11	8.05±0.21	4.96±0.31

D. Book-of-Thoth 生成提示词

本节介绍用于构建中期训练通用数据集*Book-of-Thoth*的详细信息。对于通过 KernelSynth 方法合成的多样化时间序列数据，我们设计了一组精心编写的提示词，以从 GPT-5.2 中 eliciting 高质量且多样化的时间序列-文本对。针对结构化和非结构化文本，我们分别设计了两种版本的提示词，用于生成详细文本和简短文本。这显著提升了生成数据的多样性。提示方案和示例如图9 和图10所示。

E. KnoTS 详情

KnoTS 旨在更好地评估大语言模型在知识密集型场景下的时间序列推理和决策能力。为评估现有时间序列基准的语言复杂度和多样性，我们进行了全面的词汇分析，结果见表4。词汇丰富度使用类符形符比（TTR）量化，结构多样性使用 N元语法熵（H_n）量化：

$$TTR = \frac{|V|}{N}, \quad H_n = -\sum_{w \in V_n} P(w) log P(w) \tag{4}$$

具体而言，TTR是唯一词汇量 $|V|$与标记总数 N的比值，而 H_n则根据集合 V_n中每个唯一 n元语法 w 的概率 $P(w)$计算得出。我们对均匀采样子集进行分析后发现，所提出的 KnoTS 基准测试与先前数据集之间存在显著差距。KnoTS 的类型-标记比明显更高，且 n元语法熵得分始终较高，从定量角度证实其词汇更丰富、语言结构更多样。这一优势源于我们的构建方法，它摒弃了其他数据集中常见的固定模板。

表 *4.* 不同基准测试中时间序列中心问题的多样性与复杂度分析。

数据集	类型-标记比 ↑	1-熵 ↑	2-熵 ↑	3-实体 ↑	平均长度
KNOTS（我们的方法）	**0.2109**	**9.69**	**13.02**	**13.67**	169.8
中期训练（2025年）	0.1131	7.54	10.69	12.00	77.5
时序-问答（2025年）	0.0877	7.35	8.90	9.25	49.7
CHATTIME（2025年A）	0.0098	5.56	6.86	7.40	90.4
C聊天时间序列（2024）	0.0309	7.16	8.89	9.45	293.7
TS-INSIGHTS （2025年）	0.0118	4.81	5.05	5.10	37.0

此外，我们在图11、图 12和图 13中展示了两个具有代表性的推理示例和一个决策示例。这些案例有两个共同的关键特征：它们以领域背景知识为坚实基础，并且要求模型根据具体的时间序列证据进行推理和行动。因此，KnoTS 不仅能够评估时间序列理解，还能检验引入时间序列是否会损害模型的一般推理和决策能力。

F. 示例展示

本节展示三个完整的模型推理示例，如图 14、15 和 16 所示。对于每个基准测试，我们选取一个有代表性的问题，将 Thoth 的推理结果与基线大语言模型进行比较。

G. 完整结果

ChatTime基准测试将输入序列分为四种长度：64、128、256 和 512。然而，由于在 n-shot 上下文学习设置下，长度为 512 的序列会对大语言模型的输入容量构成重大挑战，因此我们将评估范围限定为长度 64、128 和 256。完整结果详见表格5。Thoth 的两个参数版本在所有测试输入长度上均表现稳健。

表*5*。ChatTime基准测试的完整结果（Wang 等人，2025a）。四项时间序列模式识别任务采用三个固定的时间序列长度：64、128 和 256。平均值表示这三种长度下的平均结果。

模型	趋势				波动性				季节性				异常值			
	64	128	256	*Avg*	64	128	256	*Avg*	64	128	256	*Avg*	64	128	256	*Avg*
专有模型																
Gemini-3-Flash-Preview	0.970	0.800	0.840	0.870	**0.880**	**0.860**	**0.920**	**0.887**	0.550	0.550	0.460	0.520	0.700	**0.800**	0.740	0.747
GPT-4o-mini	0.890	0.790	0.870	0.850	0.380	0.500	0.420	0.433	0.540	0.370	0.370	0.427	0.660	0.660	0.700	0.673
Grok-4.1-Fast	0.680	0.700	0.560	0.647	0.460	0.380	0.310	0.383	0.590	**0.730**	0.540	**0.620**	0.700	0.750	**0.790**	0.747
开源大语言模型																
Qwen3-235B-A22B-Instruct	0.940	0.820	0.840	0.867	0.500	0.560	0.480	0.513	0.550	0.350	0.440	0.447	0.710	0.650	0.640	0.667
Qwen3-30B-A3B-Instruct	0.840	0.700	0.800	0.780	0.490	0.480	0.430	0.467	0.400	0.310	0.300	0.337	0.530	0.450	0.350	0.443
Deepseek-R1-32B	0.540	0.670	0.390	0.533	0.530	0.460	0.360	0.450	0.510	0.350	0.480	0.447	0.720	0.640	0.510	0.623
Mistral-Small-24B-Instruct	0.860	0.870	0.820	0.850	0.530	0.550	0.460	0.513	0.380	0.340	0.350	0.357	0.690	0.580	0.610	0.627
Llama-3.1-8B-Instruct	0.800	0.410	0.290	0.500	0.360	0.380	0.310	0.350	0.250	0.400	0.300	0.317	0.360	0.410	0.380	0.383
Qwen3-8B	0.380	0.050	0.060	0.163	0.310	0.340	0.010	0.220	0.270	0.020	0.010	0.100	0.710	0.610	0.020	0.447
开源视觉语言模型																
Qwen3-VL-30B-A3B-Instruct	0.920	0.810	0.850	0.860	0.230	0.510	0.640	0.460	0.440	0.250	0.360	0.350	0.480	0.490	0.440	0.470
Mistral-Small-3.2-24B-Instruct	0.460	0.500	0.330	0.430	0.410	0.380	0.330	0.373	0.530	0.620	**0.640**	0.597	0.610	0.570	0.510	0.563
任务特定时间序列语言模型																
ChatTime-7B-Chat	0.990	0.980	0.970	0.980	0.820	0.910	0.880	0.870	0.760	0.720	0.710	0.730	0.980	0.990	0.920	0.963
Time-MQA(LLaMA3-8B)	0.720	0.580	0.660	0.653	0.410	0.500	0.390	0.433	0.330	0.230	0.340	0.300	0.480	0.360	0.340	0.393
Time-MQA(Qwen2.5-7B)	0.270	0.390	0.230	0.297	0.350	0.410	0.350	0.370	0.340	0.400	0.290	0.343	0.350	0.370	0.450	0.390
Ours																
Thoth-8B	**0.980**	**0.970**	**0.970**	**0.973**	0.770	0.770	0.710	0.750	**0.610**	0.580	0.620	0.603	0.780	0.760	0.720	**0.753**
Thoth-30B-A3B	**0.980**	0.920	**0.970**	0.957	0.770	0.790	0.750	0.770	0.600	0.520	0.600	0.573	**0.820**	0.720	0.680	0.740

结构化提示（详细）

请扮演一位资深数据科学家，对提供的时间序列进行高密度技术分析。输入：数据={ts}，长度={len(ts)}

O输出要求：
- 风格：专业、分析性强且简洁。不使用对话式赘语。
- 目标：重点关注结构属性和信号动态。

1 描述性摘要

- 统计信息：[长度：数值 | 起始值：V1 | 结束值：V2 | 范围：[最小值、最大值] | 均值/标准差：数值]
- 平稳性：评估均值和方差是否随时间保持不变（平稳或非平稳）。

2 趋势与动量

- 全局趋势：总体方向、斜率强度和线性度。
- 局部动态：识别加速、减速或平台期等特定区段。
- 结构突变：注意任何永久性水平位移或机制变化。

3 模式与噪声

- 季节性/周期性：识别周期性成分或重复出现的频率。
- 自相关：描述自相似性和滞后依赖的程度。
- 波动性：刻画噪声水平，以及波动性是恒定的还是呈聚集状态（异方差性）。

4 综合分析与异常值

- 异常：定位特定的点离群值或瞬时尖峰。
- 技术判断：用一段话综合概述数据的生成过程及可预测性。采用电报式专业文风：省略不必要的主语和助动词，同时保持语法清晰。

结构化提示（简要）

请扮演一位资深数据科学家。仅从给定时间序列中提取有证据支持的观测结果。输入：数据={ts}，长度={len(ts)}

输出要求：- 总输出必须恰好包含4行：1）一行统计信息（格式见下文）2–4）恰好3个项目符号，每个以“- ”开头 - 统计信息行格式（必须完全匹配）：Stats: [Length: Value | Start: V1 | End: V2 | Range: [Min, Max]]

证据项目符号要求（第2–4行）：- 内容：仅限证据。不得提供总体判断、诊断、建议或可预测性陈述。- 每个项目符号都必须具体且可从序列中证伪：- 指明具体位置（例如“早期/中期/后期区段”“接近起始/结束处”“最大峰值/谷值附近”）。- 描述可观测特征，例如：单调漂移、斜率变化、水平位移、周期性重复、振幅调制、波动聚集、突发尖峰/骤降、持续平台期、方差变化。- 一致性约束：- 各项目之间不得相互矛盾。- 避免重复；每项应涵盖不同维度（例如，一项描述趋势 / 状态，一项描述周期性 / 自相关，一项描述波动性/异常）。- 文风：专业、简洁、电报式；不使用闲聊式填充语。- 禁止：- 不得提出无法从序列中直接看出的因果关系。- 不得使用暗示最终判断的词语（例如“因此”“总体而言”“表明”“意味着”“可预测/不可预测”）。

仅返回这 4 行。

图 9. 用于构建 *Book-of-Thoth* 结构化数据的精心设计提示词示例。

非结构化提示（特征强调）

请扮演一位专业的数据科学家。请分析以下时间序列数据，重点完成两个目标：首先，简洁地全面概述关键观测值；其次，着重指出最显著、最具代表性的特征，即能够将此序列与其他序列区分开来的特征。

定义性特征可能包括（但不限于）明显的单调趋势、鲜明的季节性/周期性、突发异常值/异常、极端的波动性/稳定性、尖锐的峰值/谷值、持续平台期，或一致的变化率——请聚焦于让该时间序列具有独特辨识度的特征。

为确保多样性，请采用不同的描述视角、术语和句式。避免重复措辞或过于专业的术语（表达应通俗易懂，同时准确）。

输入：数据={ts}，长度={len(ts)}。

请生成一段连贯完整的描述（50–150 个词），将整体概述与对最显著特征的明确强调结合起来。

非结构化提示（简要总结）

请以资深数据科学家的身份，为给定时间序列给出一句极度精简的技术判断。

输入：数据={ts}，长度={len(ts)}

输出要求：
- 格式：必须恰好为一句话。不得使用项目符号或换行。- 风格：专业、信息密集、简洁；避免对话式填充语。- 内容必须在一句话中包含以下三个部分：1）生成过程假设：指出主导结构（例如趋势／季节性／状态切换／均值回归／随机游走／异方差噪声）。2）可预测性评估：说明可预测性水平（高／中／低），并简要结合结构说明理由。3）风险点：指出预测的主要失效模式（例如结构突变、波动聚集、异常值、非平稳性、弱自相关）。

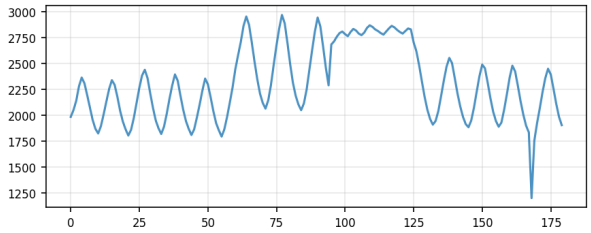
- 限制条件：
- 除必要情况外，不要输出数字；优先使用定性描述。- 不要提及提示词、“输入”或“输出要求”一词。- 不要同时提出多个相互冲突的假设；应选定最合理的主导机制。

仅返回一句话的结论。

*Fi*图 *10.* 用于构建*Book-of-Tho*非结构化数据的精心设计提示词示例 *th.*

主题：“气压泵”效应：地表大气压变化引起的土壤气体浓度波动。

情景描述：一个野外团队正在半干旱地点的生物活性土壤中，使用原位气体探针监测固定深度处的土壤二氧化碳。对于这类数据集，一条重要的解读规则是“气压泵”效应：地表大气压快速骤降时，会通过平流将深层富含二氧化碳的土壤气体向上抽吸至探针处（通常会有短暂滞后）；而气压快速上升时，则会将大气向下推入，使探针处的二氧化碳浓度被稀释。另一个关键的领域细节是，在一场透雨（或任何地表润湿事件）之后，水可能会暂时堵塞孔隙，并
降低气体渗透率，从而抑制气压驱动的平流交换；在此期间，二氧化碳对气压驱动的响应可能减弱，并且浓度可能持续偏高，因为扩散占主导地位，且逸气受到抑制。



问题：根据情景中的领域规则，哪一系列大气／地表过程最有可能导致了这条记录中的主要变化？而在通常由气压驱动的行为似乎减弱或发生改变的时段，什么现实机制可以解释这一现象？

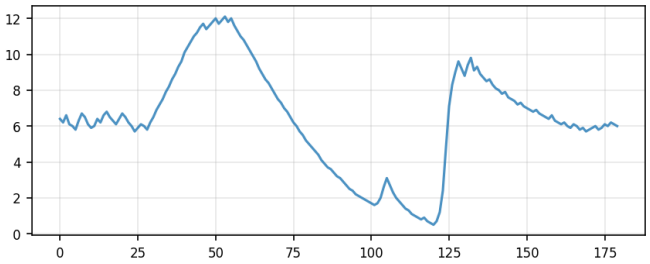
回答：结论：这段记录最符合由经过的气压系统驱动的气压泵送（气压下降时二氧化碳浓度上升，气压上升时则下降），并伴有一次暂时的地表湿润／渗透性降低事件，抑制了平流交换；一次短暂而极端的稀释事件，可能反映了大气快速增压／注入（强烈的气压上升锋面或探头扰动）。\n\n证据（模式与机理的关联）：\n- 二氧化碳浓度反复出现准周期性波动，且幅度不规则，这与天气尺度气压波动的预期响应相符（气压下降时发生平流逸气，气压上升时则发生稀释）。\n- 一段持续时间较长、浓度相对较高且较为稳定的时期，表明通风减弱而非泵送增强；这与情境中的线索一致：降雨／湿润可能封闭孔隙，使二氧化碳与气压驱动脱耦。\n- 一次突发的、异常低的浓度，最符合短暂发生的强烈向下空气注入／稀释（气压快速上升），或短暂的采样／探头吹扫事件；这会暂时以大气取代土壤气体。

图*11.* 完整的 KnoTS 基准测试中推理问题的示例。这是主文档中图4 的完整版本 t.

主题：湖水溶解氧含量的季节性周期，展示了夏季的分层和秋季的翻转。

情境描述：一个湖沼学团队在一座深水温带湖泊的沉积物上方不远处部署了一个溶解氧（DO）传感器，并让它持续工作，历经开阔水面季节直至寒冷季节。要解读这一深度的 DO 浓度，需要了解两个关键物理过程。首先，湖泊一旦形成热分层，深水层（湖底层）便基本与大气复氧和光合作用产氧隔绝，而微生物分解和沉积物耗氧量仍在持续消耗氧气；这可能导致底层水缺氧，甚至近无氧。其次，秋季表层水冷却时，密度差异减弱，湖泊便可能发生翻转

（翻转），使富氧表层水迅速向下混合，并为深水层重新充氧。此外，较冷的水能够容纳更多氧气，因此混合后深水中的溶解氧在凉爽季节可能高于温暖时期，即使生物活动没有变化也是如此。



问题：根据上述湖泊过程，推断：(i) 湖泊最可能在何时处于强热分层状态，何时处于充分混合状态；(ii) 深水是否出现了具有生态学意义的缺氧/无氧现象？哪种物理机制最能解释主要的状态变化？又有什么次要的短暂过程可以解释任何短暂的中断？

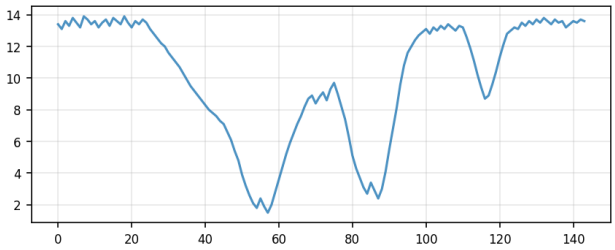
回答：该级数显示，水体从充分混合、含氧的时期转入夏季分层阶段，湖底层氧逐渐耗竭并趋近无氧状态，随后秋季翻转使底层水迅速重新充氧。\n\n结合领域知识的证据：\n- 溶解氧从高值持续、平稳地下降至~0.5 mg/L（大致对应索引 ~56–120），符合分层期间湖底层与表层隔绝、沉积物和微生物持续消耗氧气的情况；达到 <2 mg/L 表明出现缺氧，而接近 ~0.5 mg/L 的数值则表明接近无氧。\n- 溶解氧从极低值骤然阶梯式回升至 ~8–10 mg/L（大致对应索引 ~121–130），这是秋季翻转的典型特征：随着水温降低，密度分层被削弱，富氧表层水向下混合。\n- 在下降过程中出现的短暂回升（大致对应索引 ~103–106），符合短暂的部分混合事件（例如强风或内波），这类事件会暂时将一些氧气带入深层，但不会彻底结束分层。

图*1*KnoTS 基准测试中的推理问题完整示例。

主题：卫星互联网雨衰：局部降水强度（毫米/小时）与地面终端信噪比之间的负相关关系。

场景描述：远程地面终端通过容易受到雨衰影响的高频卫星链路，为现场作业站点提供主要连接。网络控制器必须根据观测到的链路质量变化，决定何时 (1) 启用吞吐量较低但稳健性很高的调制编码配置，(2) 临时改道延迟容忍型流量并暂停非必要的大批量传输，以保障关键服务，以及 (3) 恢复为高吞吐量配置。

天气单体经过该区域时，短暂的下降可以忽略，但持续恶化表明存在降水引起的衰减；若不采取缓解措施，将导致丢包和会话中断。



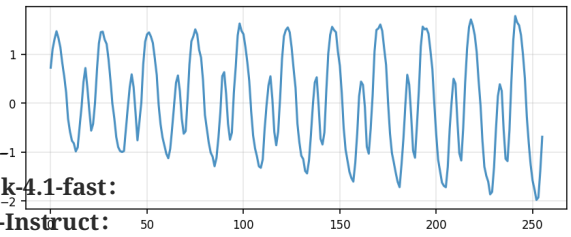
问题：仅根据提供的数值，控制器应在什么时点 (a) 从高吞吐量配置切换到稳健的雨衰缓解配置，(b) 因严重衰减风险而进一步采取措施，暂停或改道非必要流量，以及 (c) 恢复为高吞吐量配置？请指出决策点（按索引范围），并根据观测到的模式说明理由。

回答：当信噪比持续跌破 ~8 dB 时，切换到稳健的缓解配置：这种情况从索引 42（7.8）开始，并在第一次明显衰减期间持续走低（索引 42–75 的数值大多保持在 <10，并从 ~12 明显单调下降至 ~7，随后进入持续低迷阶段）。当衰减变得严重时（信噪比在多个采样点低于 <4 dB），应暂停或改道非必要流量：第一个严重区间为索引 50–60（从 3.9 降至 ~1.5–3.6，期间仅短暂回升），第二个严重区间为索引 82–88（从 3.7 降至 ~2.4–3.0）。只有在连续多个采样点稳定恢复至 ~10 dB 以上后，才切回高吞吐量配置：恢复大约从索引 94–100 开始（10.8、11.6、12.0、12.4、12.7、12.9、13.1），并在此后一直保持在基线附近，直至索引 110，表明降雨衰减事件已经结束。索引 114–118 处的后续低谷最低约为 8.7–9.6，持续时间较短、程度也较轻，不足以启动全面暂停或改道，只需继续监测（如果已经启用较温和的稳健配置，最多维持该配置）。

图*13.* KnoTS 基准测试中决策问题的完整示例。

问题：
根据以下定义：- 季节性模式：时间序列在整个期间呈现重复且可预测的波动。- 固定季节性模式：季节性波动的时间和幅度随时间保持不变。- 变化的季节性模式：季节性波动的时间或幅度随时间发生变化。请选择以下最能描述所提供时间序列的答案：

(a) 该时间序列没有明显的季节性模式。(b) 该时间序列具有固定季节性模式。(c) 该时间序列具有变化的季节性模式。仅回答 (a)、(b) 或 (c)。



真实值: (b)

模型回答：
— Gemini-3-flash：（c）— GPT-4o-mini：（c）— Grok-4.1-fast：（c）— Qwen3-235B-A22B-Instruct：（c）— Qwen3-30B-A3B-Instruct：（c）— Qwen3-VL-30B-A3B-Instruct：（c）— Thoth-30B-A3B：（b）— Thoth-8B：（b）

*Fi*图 *14.* 展示 来自不同大型语言模型的季节性识别问题 在 ChatTime基准测试中的表现.

问题：
总结这些数据点的关键统计特性：[0.14、0.24、0.04、0.18、0.52、0.69、0.81、0.77、0.8、0.67、0.33、0.31、-0.06、-0.08、-0.2、-0.16、-0.06、0.2、0.47、0.68、1.0、1.09、1.12、0.94]

真实值：该级数整体呈上升趋势，偶尔会小幅骤降。均值为正，表明总体有所上升；方差相对较低，显示其具有一定稳定性，但偶尔也会出现波动。

Gemini-3-flash：（总体得分：2）
该级数呈现出明显的周期性或波浪状模式，包含两个显著的峰值和一个谷值。其范围从最小值 -0.2 到最大值 1.12，总范围为 1.32。数据平滑变化：先升至 0.81 的峰值，随后下降至负值区间，达到 -0.2，之后回升至更高的峰值 1.12。均值约为 0.44，表明尽管该级数波动显著，但整体平均值仍为正。没有明显的异常，因为相邻数据点之间的变化相对平缓。

Qwen3-235B-A22B-Instruct：（总体得分：5）
数据点 [0.14, 0.24, 0.04, 0.18, 0.52, 0.69, 0.81, 0.77, 0.8, 0.67, 0.33, 0.31, -0.06, -0.08, -0.2, -0.16, -0.06, 0.2, 0.47, 0.68, 1.0, 1.09, 1.12, 0.94] 的关键统计特性如下：\n\n- **极差与极值**：数据范围从最小值 -0.2 到最大值 1.12，总跨度为 1.32 个单位。\n- **集中趋势**：均值（平均值）约为 0.37，计算方法是将所有数值相加后除以 24。尽管期间出现负值读数，但这表明中心值呈中等程度的正值。

总的来说，该级数表现出中等波动性，整体呈振荡行为，且在末尾呈上升趋势；均值接近 0.37，反映出早期稳定、序列中段下跌以及最后几个观测值显著急剧上升之间的不对称性。

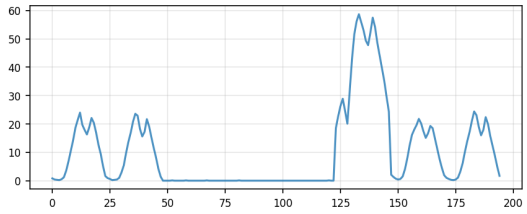
Thoth-30B-A3B：（总体得分：9）
该级数起始于较低的正值，随后跌入负值区间，再升至约 1.12 的峰值，然后趋于稳定。均值约为 0.49，中位数为 0.47，范围为 -0.2 至 1.12，表明其存在中等程度的变异，并呈现明显的上升趋势和短暂的负值下跌。

图*15*。 Time-MQA 基准中三个大语言模型对一个开放式问题的回答示例。

主题：‘数字排毒’后的反弹：在连续 3 天完全不使用设备后，接下来的 24 小时内亮屏时间大幅飙升。

情境描述：该数据集来自一项基于应用的健康研究，研究记录参与者手机内置数字健康框架（例如 iOS屏幕使用时间 / Android数字健康）中的每小时“亮屏时间”。一个重要的测量细节是：该指标只统计屏幕实际亮起且设备正被主动使用（已解锁/有交互）的分钟数；后台活动（推送通知、下载、屏幕关闭时播放音乐）不会计入亮屏时间。在这项研究中，部分参与者会参加结构化的正念静修，期间必须将手机放入

手机被放在密封锁盒中整整 72 小时；手机归还后，研究人员经常会观察到“补进度”行为，因为积压的消息和错过的更新会促使参与者长时间查看并回复消息。此外，如果手机只是关机，或记录器发生故障，研究平台通常会显示数据间断/缺失样本，而不是一长段精确的近零值。



问题：结合“亮屏时间”的记录方式和研究中的静修方案，最可能是怎样的现实事件序列和参与者行为导致了这条记录？有哪些证据支持将其解读为真实的戒断与反弹事件，而非数据采集失败？

真实值：结论：参与者很可能参加了研究中的 72 小时手机锁盒正念静修（真正没有交互使用），取回手机后立即持续进行了一场“补进度”式的刷手机活动（回复积压的消息、查看应用），之后才逐渐恢复日常节律。\n\n证据：\n- 级数中有一段很长且不间断的时段，屏幕亮起时间几乎为零，符合强制不使用手机的情况；结合场景中关于记录细节的描述，精确的近零值表明没有解锁或交互使用，而非后台活动。\n- 这段时段之后，紧接着出现了一个单日区间，每小时屏幕亮起时间大幅增加（包括异常高强度的晚间/夜间使用），符合排毒后清理积压事务、并因通知而重新参与的情况。\n- 反弹日之后，数值恢复为合理的昼夜使用模式，并呈现正常波动，这不支持传感器/记录器持续故障的解释；场景说明，真正的记录失败更常表现为缺失样本，而不是长时间精确为零。

Grok-4.1-fast：（总分：5）
结论：这条记录呈现了两个完整的正念静修周期，每个周期都包括 ~72 小时的真实手机戒断（手机放在锁箱中，有意不使用屏幕），随后出现典型的“追补”反弹，亮屏时间升高，因为参与者查看并回复积攒的消息。第一个周期跨越索引 ~49–121（73个小时点，使用量接近于零），结束时反弹达到峰值58.7；第二个周期跨越索引 ~150–222（73个点），结束时峰值为24.4。其间和之后夹杂着三个正常日周期：早晨使用量逐渐上升至午间峰值（~20–25分钟/小时），随后傍晚逐渐回落，夜间接近于零。\n\n**推理：**\n- **零值前后都重复出现相同的正常日周期（索引1–48、122–149、223–结束值）**每个周期都呈现稳定的生理规律——夜间使用量低（<1 分钟/小时），早晨上升（到午间达到20–25分钟/小时的峰值），午后下降，傍晚逐渐停止。这排除了设备故障，因为情境说明故障会导致数据中断或缺失，而不是形成有规律的日间波动。\n- **戒断阶段是真实的锁箱时段（而非关机）：**连续恰好 ~72 小时（3天）数值始终接近于零（大多为0.0，偶尔为0.1，可能是短暂查看或解锁所致），符合“在密封锁箱中放置整整72小时”的方案。期间没有数据中断，说明记录仍在正常进行，只是参与者有意不使用手机（手机开着，但屏幕关闭且无人使用）。\n- **反弹阶段证实了追补行为：**零值阶段刚结束，屏幕使用时间便猛增至异常峰值（第一次为58.7分钟/小时；第二次为24.4），远高于正常的~25最大值，之后又呈现与正常周期相似的回落趋势。这与“积攒的消息和错过的更新促使参与者延长查看和回复时间”的“追补行为”完全吻合，因而有别于不会造成使用量飙升的设备故障。\n- **没有故障迹象：**连续一致的近零值缺少电量耗尽或部分记录错误所呈现的随机性和波动；反弹则体现了有意的高强度使用，而非数据异常。持续时间恰好为72小时的 ×2，也进一步印证了这是遵循方案的结果，而非巧合。\n\n这段记录很可能反映了参与者连续参加两次为期3天的静修（例如周末 + 及之后几天），非静修日则保持正常使用手机。

Thoth-30B-A3B：（总分：8）
Conclusion: 最符合这种记录：经历 72 小时的手机戒断期（正念静修锁箱）后重新使用手机，并因积压而集中查看和回复消息，随后恢复正常使用。推理：该时间序列显示出一段持续较长、数值近乎为零的时期，与手机无法使用的情况相符（没有记录屏幕开启分钟数）。这比设备关机或日志记录故障更符合实际情况，因为后者通常会产生间隔或缺失样本。零值时期一结束，指标便立即出现急剧且持续数日的升高，数值异常偏高，与领域线索相符：积累的消息和更新会促使用户在补看期间长时间查看手机。之后，该指标恢复为此前的日周期模式，这表明零值区间是一次强制中断，而非永久性变化。

图*16*。 Grok-4.1-Fast 和 Thoth-30B-A3B 在 KnoTS 基准测试中的推理问题示例。